Analyzing 2 Categorical Variables

- **You can create categorical variables...** grouping quant.
- **Categorical Variables are often shown ...** 2-way table

	Frosh	Soph	Jr	Senior	Total
Male					
Female					
Total					n

Identify:

- **row variable** gender
- **column variable** grade
- **values of the variables** gender: M or F Grade: Fr, So, Jr, Sr
- **total (n)**
- **cells** - where data is (f)
- **totals**

rows: 2
columns: 4
2x4

Example:

	Hospital A	Hostpital B	
Died	63	16	79
Survived	2037	784	2821
	2100	800	2900

① % of all ppl. A ? $\frac{2100}{2900}$

② % of ppl. that went to A died? $\frac{63}{2100}$

③ out of everyone, what % went to B and surv.? $\frac{784}{2900}$

④ of survived, what % B? $\frac{784}{2821}$

2 types of Distributions

1) Marginal Distributions

- How to make: margins (totals) ÷ n
- Looking for .. overall % of each value of var.
- ALWAYS in %
- Hospital Example:

Example:

	Hospital A	Hostpital B	totals
Died	63	16	79
Survived	2037	784	2821
totals	2100	800	n 2900

S

- Marginal Distribution for ***Survival*** Variable:

Died: 79/2900 = 2.72%

Surv: 2821/2900 = 97.28%

- ***Marginal Distribution for Hospital* Variable:**

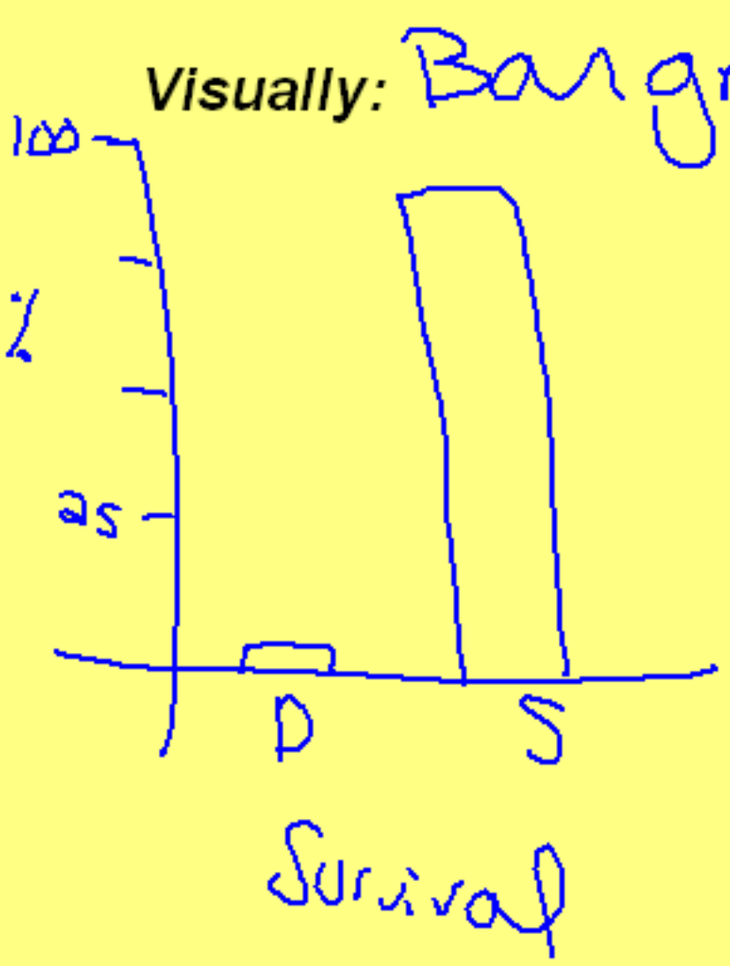
Example:

	Hospital A	Hostpital B	
Died	63	16	79
Survived	2037	784	2821
	2100	800	2900

A: 2100/2900 = 72.4%

B: 800/2900 = 27.6%

Visually: Bar graph



1) Conditional Distributions

- **Look at ...** one variable
- **Then look at ...** one value at a time
- **Break down ...** value into its parts
- **ALWAYS** in %
- **Hospital Example:**
  - **Conditional Distribution for *Survival* Variable:**

Example:

	Hospital A	Hostpital B	
Died	63	16	79
Survived	2037	784	2821
	2100	800	2900

Died

A: $63/79 = 79.7\%$

B: $16/79 = 20.3\%$

100%

Survived

A: $2037/2821 = 72.2\%$

B: $784/2821 = 27.8\%$

100%

- ***Conditional Distribution for Hospital* Variable:**

A
D: 63/2100 = 3%
S: 2037/2100 = 97%
100%

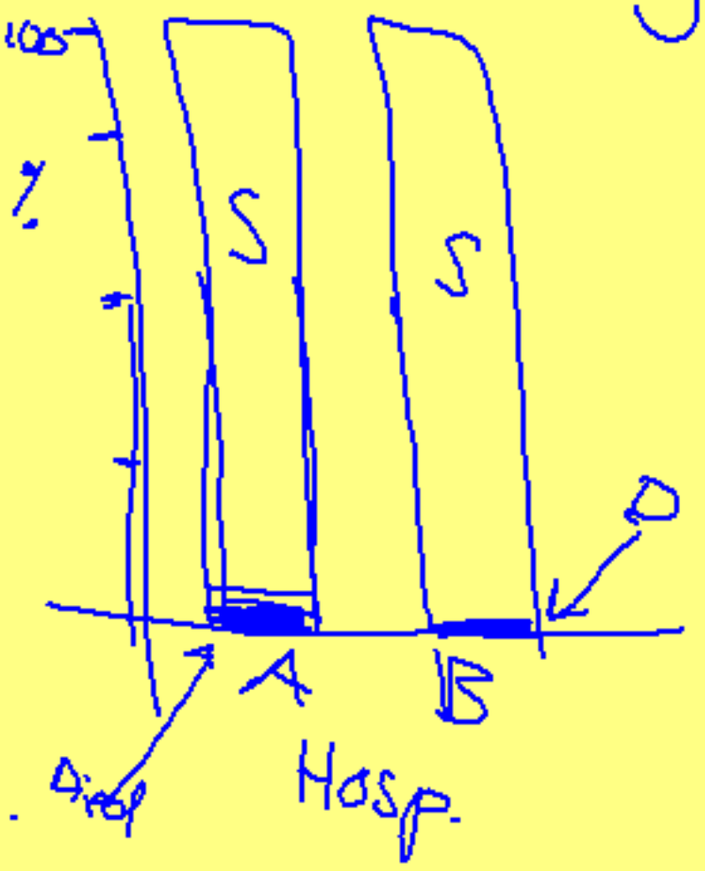
B
D: 18/800 = 2%
S: 784/800 = 98%
100%

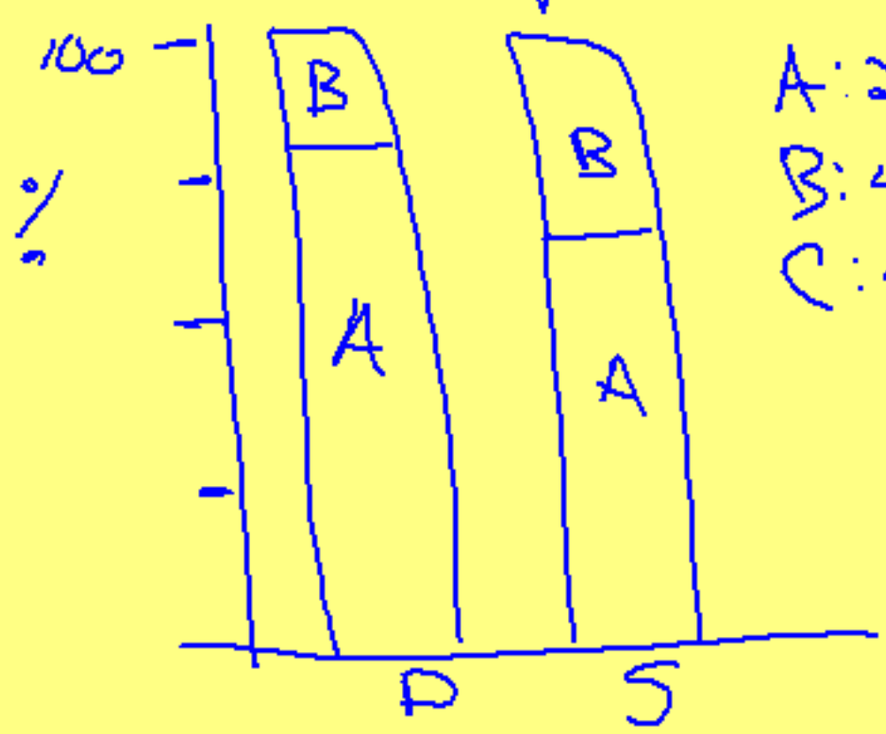
Example:

	Hospital A	Hostpital B	
Died	63	16	79
Survived	2037	784	2821
	2100	800	2900

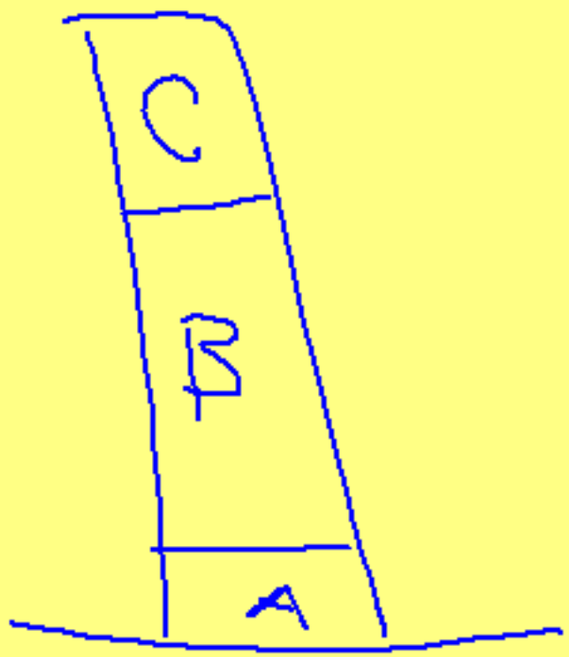
Represented Visually: Stacked/Segmented Bar Graph

- **Each bar = 100%**
- **Conditional Variable on ...** X-axis
- **Bars are ...** segmented into parts of value



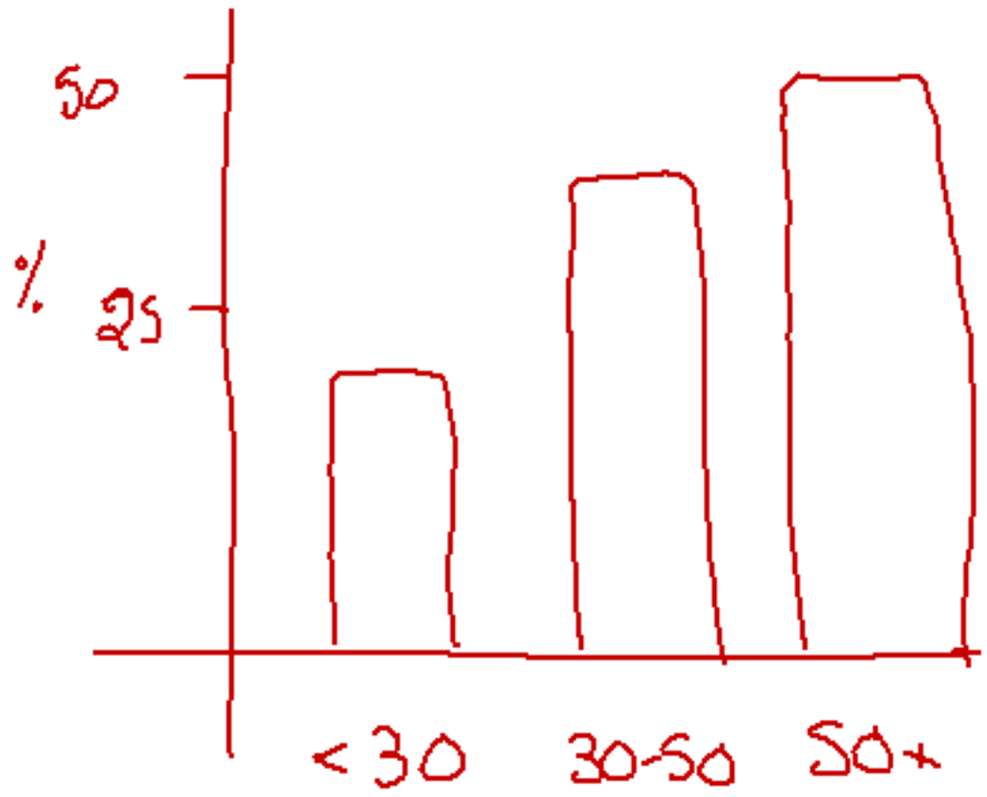


A: 20%
B: 40%
C: 40%



Marginal

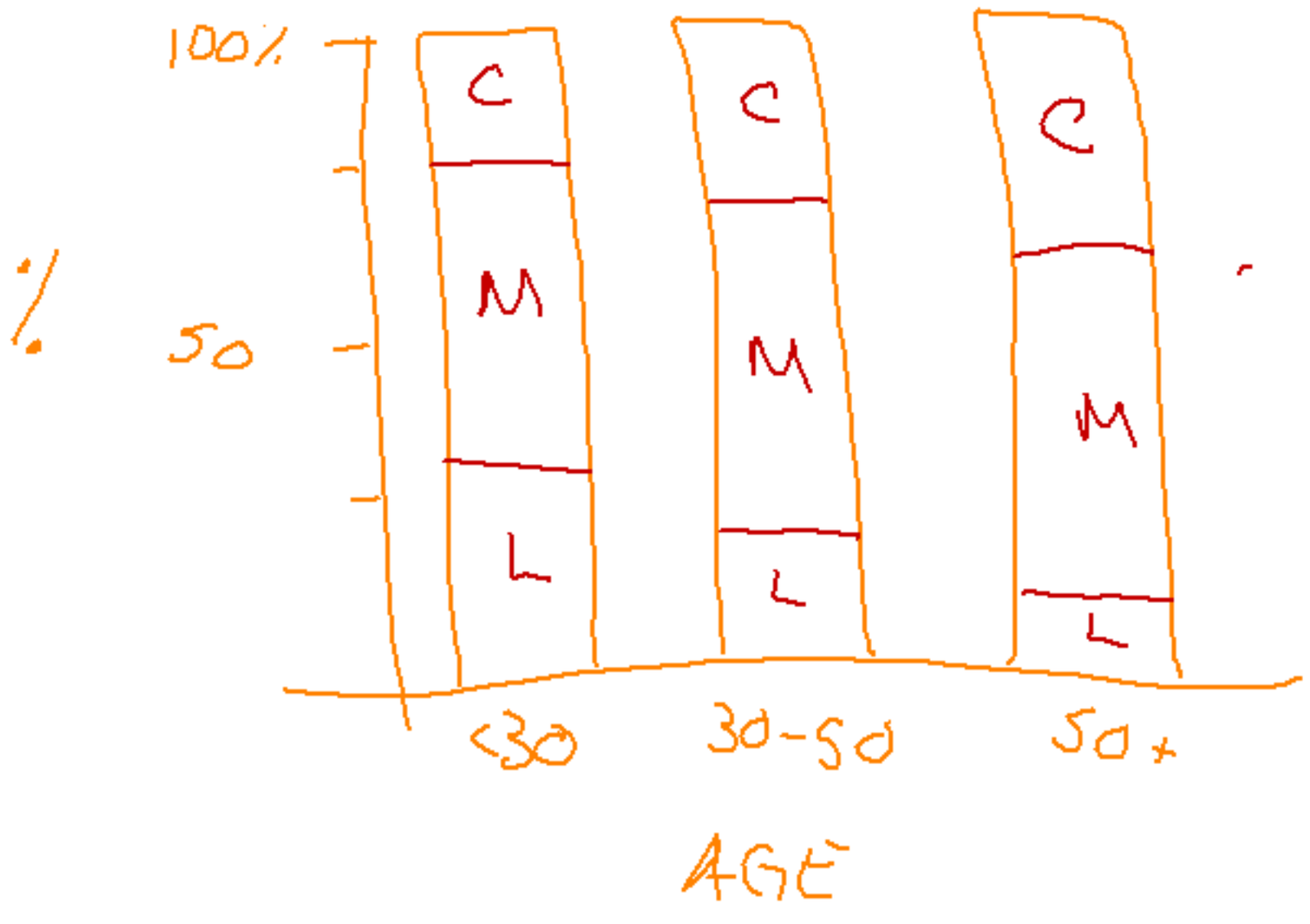
Age



Political Pref.

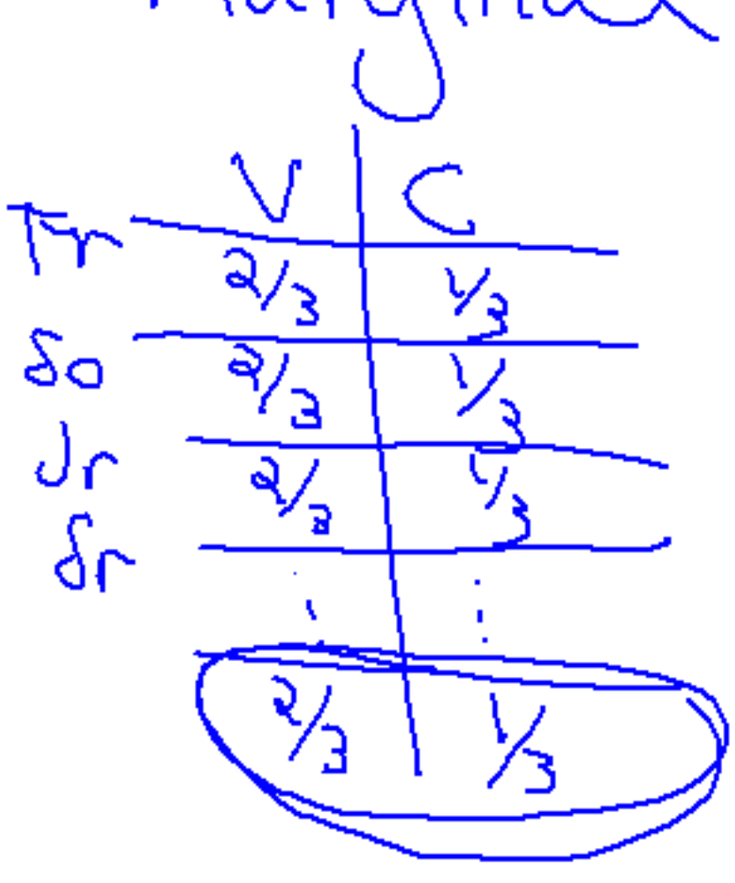
Conditional - Age

	<30	30-50	50+
L	28%	21.3%	15%
* M	47.3%	50%	48.5%
C	24.7%	28.8%	36.5%



Independence

- Independent=
  2 things are indep if
  they don't affect each other
- Marginal distr = conditional distr

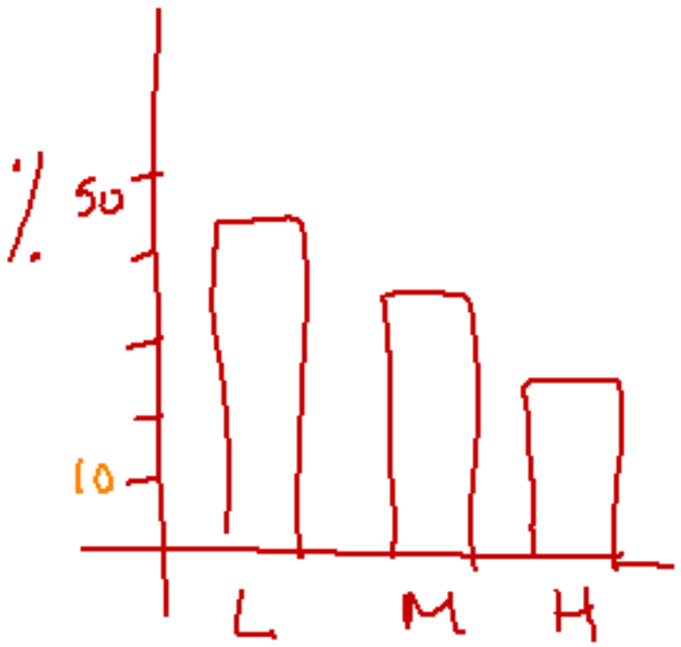


Simpson's paradox

- overall → one concl.
- Add 3rd variable,
  make a different concl.

Candi Kulp Jones

(a) Cholesterol

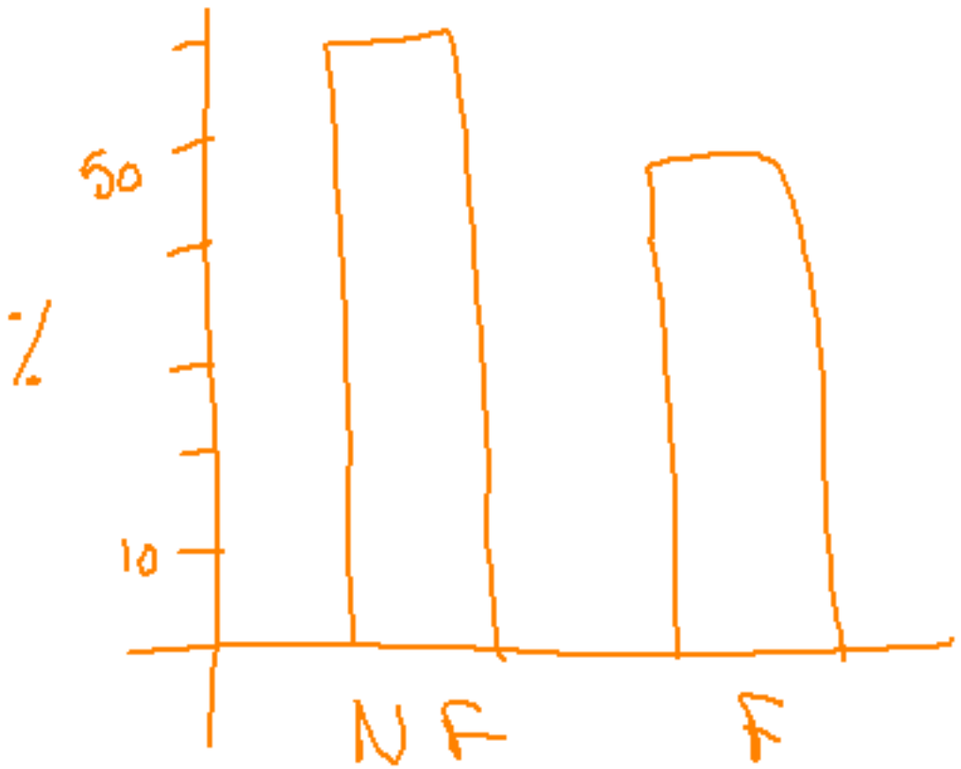


L: 42.7%

M: 33%

H: 24.1%

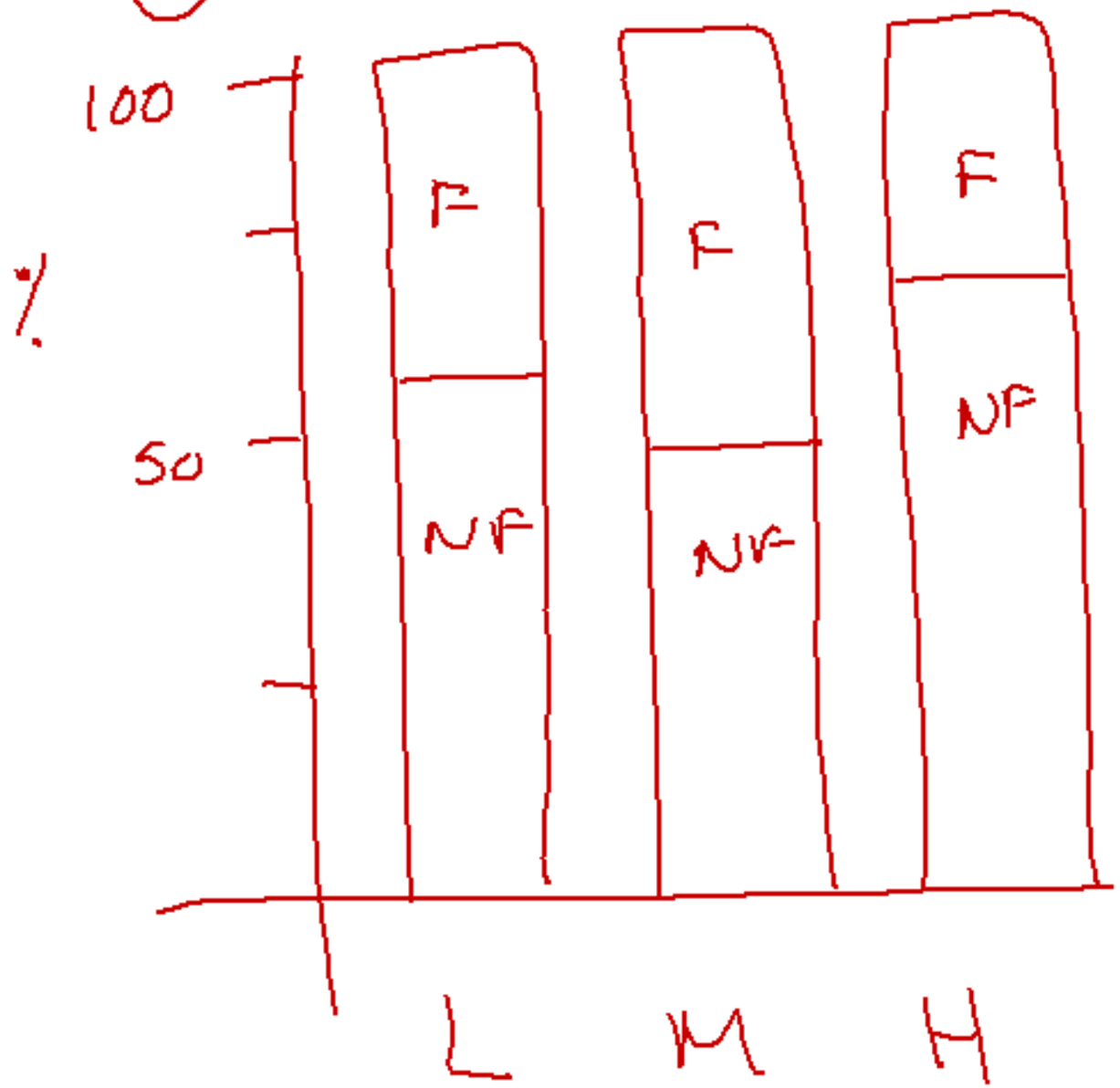
(b) Heart attack



NF: 57.14% ≈

F: 42.86%

(c) Chol. - conditional



(d) Heart attack - condit